SHARED NETWORK GOVERNANCE AND STEWARDSHIP OF DATA AND THE EXCHANGE OF DATA

A White Paper of the
INFORMATION INTEGRATION NETWORK

Draft - June 21, 2000

PURPOSE

This paper discusses the importance of shared governance and stewardship of integrated information. Shared governance and stewardship pertains to EPA, states, and other stakeholders in establishing and maintaining the environmental exchange network. It pertains to the program offices, EPA's Office of Environmental Information, and regional offices within EPA in establishing and maintaining the EPA node(s) on the network. It pertains to the data, the exchange of data, management of the databases (including registries), and maintenance of the network nodes of the national environmental information exchange network.

Background:

Currently, each program office manages or coordinates the management of data that pertains to its programs. Except for some basic facility registration information, data are not integrated across programs. Data are either publicly accessible or not accessible at all. This situation has been driven by laws that focus on only one aspect of the environment at a time, separately delegated down to individual programs offices, who, given time and budget constraints, develop regulations and systems in support of the individual programs.

As citizen interest in local and global environments grow and as access to information becomes more an more available to citizens locally (mostly via the Internet) constituents want to have a broader, overall understanding of their environment and how it affects them. We are now being asked to integrate and harmonize data sets that up to now have been collected, maintained and disseminated in very much a stovepipe fashion. We will all need to work together to make this happen, and given time and budget constraints, there may be efficiencies to centralizing much of the activity under an agreed upon set of procedures, standards and protocols.

DISCUSSION

What is Governance and Stewardship?

Governance:

In an integrating data context - **Governance** is the development and implementation of a set of rules for managing the network including data standards, protocols for exchanging data, procedures for maintaining and improving data quality, agreements for preserving security including accessibility, integrity and confidentiality and particularly, agreements on who will maintain and make accessible the authoritative copy of each of the data sets.

Shared Governance:

Network partners need to participate in forming and ensuring compliance with the procedures, protocols, standards, etc. necessary to maintain each part of the network. It is important that the network partners come to consensus on the rules they will need to follow as network stewards. There will be governance of the network across network partners, as well as governance of the network node across the offices, regions, and other components within an organization that manages that node as a partner on the network.

It is not enough just to share the data, it needs to be properly maintained (quality assured, updated, accessible, explained, etc.). Good stewardship can be ensured through trading partner agreements, performance partnership grants, and performance partnership agreements as well as other measures. Where good stewardship can not be enforced, statements about data quality and availability can be made on the sites that point to the other sites on the network.

Stewardship:
In an integrating data context - **Stewardship** is managing the data, resources or activities – from data collection, through maintenance and disposition. It is foremost the role of quality assurance, but extends to the analysts' "respectful use of data" and includes making the data available to all those (and only those, in the case of proprietary and security related issues) who are authorized to access it. Stewardship extends, also, to efficient management and effective integration with other, related data.

Stewardship is not ownership. EPA, its program offices and regions, the states - none of us actually own the data. We manage the data for taxpayers, stockholders, etc.

Shared Stewardship:
Stewardship is a shared responsibility:

- Across all organizations in the network, each organization which provides data for use by others on the network must exercise stewardship over the data, but the roles and responsibilities will vary from mere warehousing to actually ensuring and maintaining the quality and timeliness of the data. However, without good stewardship, there may exist data of suspect quality, availability and usability which forms a gap in the web of data contained in the network. Mechanisms will need to be in place to coordinate and resolve problems with the flow and sharing of the data, costs vs. value added, data quality, etc.

- Within each organization on the network, stewardship is also a corporate or agency-wide responsibility. EPA, for instance, will need to work out stewardship of the following aspects of our network:

    - Data content, integrity, and quality - by data set, data table, field and record
    - Applications to collect the data
    - Applications to process the data
    - Applications to provide access to the data to the public and other stakeholders
    - Maintenance of the database that houses the data
    - Maintenance of the hardware and operating software

This can be difficult when the data were formerly "owned" by several of the offices within the organization. However, shared stewardship has economic and access benefits and can be accomplished through agreed upon protocols, processes and standards as well as access controls.

Why do we need Shared Governance and Stewardship?

Data collection and management is costly and is best shared: Data is, increasingly, a major environmental management resource. Because of its value, and because of its high cost, it must be preserved and used by all who want to, and have legal authority to, access it. In the words of EPA CIO, Ed Levine, "If I know you have certain data, and I know how to get access to it, and it's reliable, good quality data, then I can depend on your data without having to collect it or maintain it, too!"

Data quality improvements are to be shared: In addition, improvements and corrections to the data need to be shared so that everyone benefits and can rely on the shared data source.

Data needs to be where you expect it, when you expect it, of known quality, source, etc.: Using an agreed upon set of data standards, formats, metadata, etc. makes finding the appropriate data more efficient and using the data correctly more feasible. It allows consistent and reliable transfer of data using understood standards and via compatible mechanisms.

What Kinds of Stewardship are Needed for I-3?

Stewardship is a far-reaching concept. At a high level, the network as a whole requires stewardship both by the states and EPA. Each partner will be the steward organization for its own node on the network, making sure it is functioning properly and that the data are available through it within the jointly agreed upon terms.

Each organization that is exchanging data on the network is responsible for ensuring that its data are transmitted and received in the agreed upon format and timetable, that the integrity of the data are intact, and that, in the case of confidential or other sensitive data, the data have not been intercepted. Hence, there is a need for stewards in each organization.

Individuals within the partner organizations (e.g., EPA, each state) will need to be responsible for making sure the hardware on which the data reside, and the software that secures and serves up the data are all working properly. This includes operating software, database software, applications software, etc. Within EPA, there is a need to determine how to assign stewardship responsibilities. For I-3 purposes, we may find that stewardship of the data (and tables within the databases) as well as applications to collect data should be decentralized, while stewardship of the database engines, hardware and operating software as well as applications to access the data publicly or across the agency should be centralized.

There is also the concept of a custodian who merely warehouses a copy of the data for convenience of access without any effort to improve the quality of the data or participate in governance.

Registries (See the white paper "The Proposed Use of Registries in Information Integration" for details) on the network must have shared (corporate) stewardship across the relevant constituencies if they are to be reliable and authoritative sources for commonly used facility, corporate, industrial sector, place, chemical, etc. data.[1]

Programmatic or State Data Linked to (but not actually in) Registries:
Registries will link to data sets that are not actually part of the registries. It is essential that the links (e.g., EPA facility ID or program system facility ID - whatever is the agreed upon number, EPA chemical ID or CAS number - whatever is the agreed upon number) remain intact. However, the quality of the other data within their program or state records is entirely under their control. Stewardship for that data and how any data set on the network should communicate the quality and timeliness of the data in that data set (e.g., meta data about the data set or extra fields indicating when the record was last updated, etc.) would need to be defined in an exchange format and agreed upon in a trading partner agreement. This is especially important for secondary users of their data and that meta data need to be prepared and shared to support that type of use.

Transition to stewardship of registry data:
The first registry to be established will be the regulated facility registry (FRS), maintained centrally by EPA with the support of the stewardship network. Over time, EPA (and perhaps other entities) will add additional registries for place, chemical and other substances, and expand the facility registry to include other entities that are not part of the facility registry (e.g., corporations that do not have facilities in the facility registry). As each registry is added, it should become the authoritative source of that information and be integrated with the other registries and network databases as appropriate. For more information, see the "The Proposed Use of Registries in Information Integration" white paper.

Transition to stewardship of non-registry data:
Ideally, the vision is to move toward a network where each partner on the network maintains its own data on its own server in a database that is compatible with the rest of the network and its applications. Thus, except for the registries, states and program offices could maintain their own data locally and outside users of the network could access the information through applications that would reach across the network. However, because not all network stewards are ready to provide this sort of access in a secure manner, there will clearly need to be a transition period where EPA centrally maintains accessible copies of non-registry programmatic and state data sets and acts as a custodian (without changing the data) to simplify and speed up access. For example, in the short term we may use data warehousing, whereas in the longer term we would move

---

[1]The roles for registry data stewardship described in the *Facility Registry System: Data Steward Manual - May 12, 2000 Draft*, pages 4 through 8, could easily be adapted to cover other registries that have been proposed in the "The Proposed Use of Registries in Information Integration" white paper (mostly by changing "FRS" to "EPA registries" and "FLA" to "EPA registry tools"). The roles include a Data Stewardship Manager (for the Agency as a whole), EPA Program Data Stewardship Managers, Regional Data Stewardship Coordinators, Regional Data Stewards, and Participating State Data Stewards.

toward a decentralized access ("come and get it") approach, or more likely, some combination of the two over the exchange network.

CONCLUSION/RECOMMENDATION:

Stewardship and governance are shared responsibilities. All of us involved in the environmental exchange network will need to develop and implement data standards and rules of operation that support the overall goals of the network, or it will falter.

Some data sets may originate from numerous sources, but in order for them to be fully integrated, will need to be managed centrally based on mutually agreed upon data and procedural standards. OEI within EPA may take on that role for certain data sets, but other program offices, certain states or other stakeholders may be the appropriate organizations to take on that role for other data sets.

Although each organization using the network will need to sustain a certain level of trust about the other parts, verification methods may be necessary to ensure that all data sets on the network continue to be maintain an acceptable level of quality, availability and security and metadata.

A successful I-3 network will require an associated stewardship network. We recommend that the I-3 stewardship network be expanded from the current Facility Registry System (FRS) effort to cover other registries, and that non-registry data stewarded by States and program offices be included in that stewardship network, although to a lesser extent.

APPENDIX:

Data vs. Database vs. Application vs. Network Node Stewardship:
In addition to data stewardship is the concept that there is hardware and software that serve up that data to the network. The data stewards may do a fantastic job of cleaning and maintaining the data, but if the Internet server is down, it will be inaccessible. Also, the data may be in great shape, and the server may be working just fine, but the applications that provide access to the data are not user friendly or incompatible with other applications on the network.

What are the right formats in which the data should be stored for access? Most of EPA's data are stored in Oracle databases, but XML appears to be the format for exchange of data over the Web. However, Oracle uses relational tables, whereas XML uses a hierarchical structure. Also, the format for geographic data may be different from that format for other data sets. Clearly, database standards as well as data standards are part of the solution.

In addition, there will be various types of users on the network, some of which will have access to all the data, and some of which will only have access to public data. The **database steward** will need security controls at the database level to control access by various types of users.

In order for the database security to work correctly, there will also need to be appropriate security at the server level. A **network node steward** will need to make sure these security measures are in place as well as making sure the server is up and running and connected to the Internet.

Once the data is in (a) standard format(s) on Internet servers with the appropriate security controls, there will need to be applications that give users access to that data. To the extent that the access applications are on the network node that provides the data, there will need to be an **application steward** to keep that application up and running, up-to-date, etc. and compatible with the rest of the network. To the extent that the access applications are on the user side, they are the responsibility of the user and there would not need to be an application steward.

Example: Facility Registry System (FRS):

Facility data are collected for the following programs: TRI, RMP, RCRIS and BRS (RCRAInfo), CERCLA, PCS, SDWIS, AIRS, NCDB.
In many cases the collections come through the regions; and in many cases, the states.

FRS data stewards need to include representation from each of those program offices as well as regions and states. Note that program offices that do not collect that type of data do not have obligations for that registry (e.g., OPP).

For the data stewardship to work, users on the network need to:

- Know where to find what data - so there needs to be a place from which to access all registries that is widely advertised,
- Be able to reliably access the data - so there need to be set hours of operation (7 days a week, 24 hours a day or just regular business hours?), as well as servers that can handle the level of traffic expected, and clarity on what portion of a data stewards data each user has access to (states may be able to access more than a member of the general public)
- Know the meaning, quality and currency of the data - so there need to be data standards and meta data that include definitions of each data element, interpretive data, measures of data quality that are in turn explained, and information on when the record(s) was(were) last updated, as well as procedures for populating, updating, and accessing data sets.